United States Military Academy
West Point, New York 10996

A Process to Improve the Efficiency of Stop Decisions in Group Experimental Testing

OPERATIONS RESEARCH CENTER OF EXCELLENCE
TECHNICAL REPORT #DSE-TR-03-05
DTIC #: ADA419601

Lead Analyst
LTC Andrew Glen,PhD
Assistant Professor, Department of Mathematical Sciences

Senior Investigator
Bobbie Leon Foote, Ph.D.
Professor, Department of Systems Engineering

Directed by
Lieutenant Colonel Michael J. Kwinn, Jr., Ph.D.
Associate Professor and Director, Operations Research Center of Excellence

Approved by
Colonel William K. Klimack, Ph.D.
Associate Professor and Acting Head, Department of Systems Engineering

September 2003

The Operations Research Center of Excellence is supported by the
Assistant secretary of the Army (Financial Management & Comptroller)

A process to Improve the Efficiency of Stop Decisions in Group Experimental Testing.

Lead Analyst

LTC Andrew Glen, PhD

Assistant Professor, Department of Mathematical Sciences

Senior Investigator

Bobbie Leon Foote, Ph.D.

Professor, Department of Systems Engineering

OPERATIONS RESEARCH CENTER OF EXCELLENCE
TECHNICAL REPORT #DSE-TR-03-05
DTIC #: ADA419601

Directed by

Lieutenant Colonel Michael J. Kwinn, Jr., Ph.D.

Associate Professor and Director, Operations Research Center of Excellence

Approved by

Colonel William K. Klimack, Ph.D.

Associate Professor and Acting Head, Department of Systems Engineering

September 2003

The Operations Research Center of Excellence is supported by the Assistant Secretary of the Army
(Financial Management & Comptroller)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE 02/04	**2. REPORT TYPE** Technical report	**3. DATES COVERED** *(From - To)* 01/03-02/04

4. TITLE AND SUBTITLE
A process to improve the efficiency of stop decisions in group Experimental testing

5a. CONTRACT NUMBER
DSE-R-03-05

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
LTC Andrew Glen

Bobbie Leon Foote,PhD

5d. PROJECT NUMBER
DSE-R-03-05

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

ORCEN
SYSTEMS ENGINEERING
USMA, WEST POINT, NY, 10996

8. PERFORMING ORGANIZATION REPORT NUMBER
DSE-TR-03-05

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Warrior/ Operational Test Command
David Laack
LaackDavid@otc.army.mil

10. SPONSOR/MONITOR'S ACRONYM(S)
Air Warrior (OTC)
USAOTC,AVTD,Cbt Test Div

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
PUBLIC DISTRIBUTION

13. SUPPLEMENTARY NOTES
SOFTWARE IS AVAILABLE TO PERFORM TEST FROM THE AUTHORS

14. ABSTRACT

A process has been designed to stop early after r items have failed when n items are on test at the same time. The prior distribution can be any of a number of well known distributions. The power of the decision sometimes exceeds a full sample and is normally very high. Large savings in test facility capacity, item destruction and test personnel time can be achieved.

15. SUBJECT TERMS
CENSORED TEST, ORDER STATISTICS, TYPE II CENSORING

16. SECURITY CLASSIFICATION OF: **unclassified**			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** LTC Mike Kwinn
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**		27	**19b. TELEPHONE NUMBER** *(include area code)* 845 938 6493

Abstract

This report documents work done for Air Warrior to make simultaneous testing more efficient. If N items are on test, then after r items a decision can be made as failure time are reported. We show that at times r<.2N. This allows items to be reused if the test is non-destructive. This test is ostensibly for electronic items, but is versatile and offers great advantages for Army research testing the effectiveness of medications, vaccines, and other new medications. It is possible to reject a drug or accept it with high power very quickly. It can also be used to check if a deployment was especially dangerous and give warning to pay special attentions to veterans of certain theaters.

About the Author(s)

The authors are Professors at USMA, WEST POINT, NY, 10996.

Acknowledgements

We thank Air Warrior for funding this work, Capt. Mary Lou Hall for her steadfast support, for the encouragement of Col. William Klimack, and the editing, suggestions and support of Professor Larry Leemis, William and Mary University.

Table of Contents

Goodness-of-Fit for Sequentially Censored Lifetests

LTC Andrew G. Glen, Ph.D.
Department of Mathematical Sciences
United States Military Academy
West Point, New York
aa1275@usma.edu

Bobbie Leon Foote, Ph.D.
Department of Systems Engineering
United States Military Academy
West Point, New York
fb9690@usma.edu

June 12, 2003

Abstract

We propose a methodology for gaining statistical inference on censored samples, especially during the actual conduct of the lifetest experiment, in order to reduce cost and time on test while preserving reasonable levels of statistical power, and in at least one case, the methodology has increasing statistical power of a censored sample over that of a full sample. The outcome of the methodology will produce design efficiencies in lifetime testing. The method is distribution free for any fully specified continuous distribution under the null hypothesis, and produces p-values that are exact. Transforming ordered lifetest data into iid uniformly distributed data on (0,1), we use the T_n statistic, discussed in a companion paper (Glen and Foote 2003), to gain inference on mean life of systems with resulting power increases of up to 30% higher than that of the Anderson–Darling statistic. We investigate, with simulation, the power of the method as r (the number of failures currently observed) increases to n. We look specifically at null hypotheses from the exponential, normal, and gamma families of random variables. We introduce an automated tool that allows for immediate implementation of the new method using the probabilistic software package "A Probability Programming

Language" running in the Maple environment. We provide conclusions that will give insight on how to gain statistical inference with less time and materiel on test. We also show a counter-intuitive result where in certain cases, censored samples produce higher power than full samples. We investigate this counter-intuitive result more fully.

Keywords: Computational Algebra Systems, Exact Distributions, Conditional Order Statistics, Censored Lifetesting.

1 Introduction and Literature Review

In lifetesting applications, tests are designed to gain an understanding of the probabilistic properties of a component or a system of components. Often, the costs of lifetests, in both time and money, constrain the design of the experiment, limiting the number of items placed on test and the length of the test. Many times, like in pharmaceutical drug testing, the length of the experiment cannot be estimated accurately in advance, and often one is faced with un-analysed, censored data in an ongoing experiment. For such cases we propose a methodology that gives exact statistical inference on censored samples. Consider an existing component, process, or drug with an all-parameters known lifetime reliability distribution $F(x)$. Should an improved component, process, or drug come along, both producers and consumers would like to verify that the new item is better than the existing item, most often by determining if its mean lifetime has improved (whether a decrease or an increase). In the lifetesting of the new component, it would be highly desirable to stop the test when enough evidence exists to support either claim. Such censoring, commonly called Type I (stop after time t) or Type II (stop after r items fail), can produce statistical inference, however, existing methods are not widely known, nor do they have remarkable statistical power. We propose a methodology that will specifically rely on Type II censoring in the design and conduct of the lifetest. If for example, one could afford a lifetest with $n = 5$ items to fail, a certain level of statistical power could be achieved if the test continued until completion of n failures. Consider, however, an example where $n = 25$ items are placed on test with $r = 5$ as the designated censoring value. Obviously the second test would conclude more quickly, as the expected time on test would be the mean failure time of $X_{(25:5)}$, the fifth order statistic from a sample of 25 items, under the null hypothesis. Now consider a slightly different example, where $n = 25$ items are placed on test. Experimenters notice that after $r = 3$ failures, lifetimes seem to be substantially better than the original system. After $r = 6$ failures, they are convinced, at least anecdotally, that the new system is better. We propose a new methodology and a new test statistic that will allow for instantaneous assessment at every failure, with exact p-values, from an exact distribution of the test statistic. We rely on properties of conditional order statistic distributions to provide statistical inference for censored data that has acceptable statistical power. We also show that for the case of the Gamma distribution, given certain conditions, it is possible to achieve higher power with a censored sample than it is for a full sample, a counter-intuitive result that has warranted in depth investigation on our part. We use the test statistic T_n, presented in a companion paper (Glen and Foote 2003), which has significantly more power than the Anderson–Darling statistic, given changes in mean lifetime. The method we propose transforms censored data, via two probability integral transforms (PITs) and conditional order statistics, into an un-ordered, iid sample of uniformly distributed data on the open interval (0,1), which we abbreviate $U(0,1)$. Furthermore, the test statistic T_n, designed as a test of uniformity, enjoys significantly higher power than the A–D statistic when finding differences in the mean of the distribution of the item in question, thus higher power is generally possible by combining the censored methodology with the use of the T_n statistic. The net effect of combining the new statistic with the new methodology is an very strong advantage in assessing censored data, to include the possibility of purposefully designing lifetests with higher values of n so that the test can be censored early at a reasonable value of r, saving time, money, and items that were destroyed during the test.

Rosenblatt (1952) presents theory that transforms joint conditional statistics to ordered, uniformly distributed statistics for the censored case (we will instead transform censored data to a complete un-ordered set of uniform data). David (1981) discusses the Markov nature of conditional order statistics, and equates the conditional order statistic with the truncated order statistic, a result that we will use as part of our method. O'Reilly and Stephens (1988) use a Rosenblatt transform, then invert that transformed data to test ordered uniform data (we will test un-ordered uniform data). Hegazy and Green (1975) present work on goodness-of-fit using expected values of order statistics with approximations used for critical values. Balakrishnan, Ng, and Kannan (2002) present a test for exponentiality that is based on progressively censored data, which uses a T statistic, however this statistic and this method is unrelated to the T_n statistic and the sequentially censored data analysis that we use. Michael and Schucany (1979) also present a transformation that takes censored data and transforms it into ordered uniform data. Since Michael as well as Stephens (1974) also point out that the A–D statistic is generally more powerful than the other well-known goodness-of-fit statistics in the case when the mean has shifted, we will rely on T_n, which has even higher power in detecting shifts in the mean than A–D, as shown our earlier, companion paper (Glen and Foote, 2003).

2 Transforming the Censored Data into IID $U(0, 1)$

Let the lifetime of an existing system (also that of the null hypothesis) be distributed by the all-parameters known continuous rv X with CDF $F(x)$. Let n items be on lifetest and let the Type II censoring value be r. Recall that in a lifetest, failure data arrives in ordered fashion. The ordered lifetime data $x_{(i)}$ have CDFs from their appropriate order statistic $F_{X_{(n:i)}}(x_{(n:i)}),\ i = 1, 2, \ldots, r$, (note $X_{(n:i)}$ is abbreviated $X_{(i)}$). Now consider the conditional order statistics of the lifetest, $X_{(1)}, X_{(2)}|X_{(1)}, \ldots, X_{(r)}|X_{(r-1)}$. Theorem 2.7 from David (1981, pg. 20) explains the Markov nature of these conditional order statistics. Thus for our purposes the CDF of the i^{th} order statistic, given the $(i-1)^{\text{th}}$ data point, $F(x_{(i)}|x_{(i-1)})$, is that of the rv $X_{(n-i+1:1)}$ with support $x_{(i-1)} < x_{(i)} < 1$. David shows this is the first order statistic from a sample size $n-(i-1)$ from the parent distribution of X truncated on the left at $x_{(i-1)}$. In other words, the distribution is independent of $x_{(1)}, x_{(2)}, \ldots, x_{(i-3)}$, and $x_{(i-2)}$, and is therefore memoryless in this regard. Since each of the conditional distributions can be computed, conducting separate PITs on each data value, $F_{X_{(i)}|X_{(i-1)}}(x_{(i)}),\ i = 2, 3, \ldots, r$ will give a sample of r iid $U(0, 1)$ random variables (see Rosenblatt 1952, pg. 470) to which a uniformity test can be applied. As mentioned earlier, we use T_n, as it is better at finding changes in μ_X than A–D in many cases (Glen and Foote, 2003). The statistic T_n has the distribution of the convolution of n iid $U(0, 1)$ random variables. Therefore, the test statistic we will use is as follows:

$$T_r = \sum_{i=1}^{r} F_{X_{(i)}|X_{(i-1)}}(x_{(i)}),$$

where $F_{X_{(1)}|X_{(0)}}$ is defined to be $F_{X_{(1)}}$, and r is the size of the censored sample.

3 Implementation using APPL

The theory of the statistic is straightforward, however the implementation is made practicable only with automated probabilistic software. We implement the new method and new statistic in APPL (Glen, et. al. 2001) for a number of reasons. The software allows us to use exact distributions of the original data, the distributions of the conditional order statistics, and the distribution of the T_n statistic so that exact p-values are attainable. Additionally, the author has already calculated the PDFs of the sum of n $U(0,1)$ random variables from $n = 1$ to $n = 50$, the last PDF requiring 91 pages of ASCII text to enumerate. APPL reads these PDFs exactly and can thus compute the exact p-values. APPL allows for the use of any continuous distribution (well-known distributions as well as ad hoc) to specify the null hypothesis and conducts the necessary PITs for these distributions. We will demonstrate power of the censored and full samples using T_n and A–D statistics with data from the Normal, Exponential, and Gamma prior distributions, however we are not limited to just these distributions.

The methodology can be confusing to those not used to using conditional order statistics, thus we present more clearly the algorithm for computing the test statistic.

- Specify the null distribution of the existing (old) system, $F(x)$.

- During the lifetest experiment, note n and create the vector of r observed occurrences.

- Calculate $z_{(i)} = F(x_{(i)})$, $i = 1, 2, \ldots, r$, which is ordered uniform (not iid).

- Calculate the unordered, iid $U(0,1)$ (under the null hypothesis) $u_i = F_{Z_{(i)}|Z_{(i-1)}}(z_{(i)})$, $i = 1, 2, \ldots, r$. Note: we perform the PIT with F(x) and then conduct the conditional order statistics PIT using the uniform conditional order statistic distributions. These two methods have been shown to be equivalent (Glen, et. al., 2001), but this method is preferred as the conditional order statistics of the uniform distributions are much more tractable than conditional order statistics using the parent distribution F. Also note, we find the conditional order statistic using the truncation of the parent distribution method outlined by David (1981).

- Sum the u_i values to get the T_r statistic.
- Calculate the p-value with the appropriate tail of the T_r distribution.

The APPL code that enacts this algorithm to calculate the statistic is as follows:

```
# take the r censored values in 'data' and PIT them into the list 'Zdata'
for i from 1 to r do
  Zdata := [op(Zdata), CDF(Nulldist, data[i])];
od;
# sum the independent uniforms to for the statistic 't_stat' starting with the first failure ...
t_stat:=CDF(OrderStat(U(0, 1), n, 1), Zdata[1]);
# ... then adding up the subsequent failures until r is reached.
if (r > 1) then
  for i from 2 to r do
    t_stat := t_stat + CDF(OrderStat(Truncate(U(0,1), evalf(Zdata[i-1]), 1),
      n - (i - 1), 1), Zdata[i]);
  od;
fi;
Tr_distn := cat('T',r);
# now return the statistic, the lower tail pvalue and the upper tail pvalue
# using the APPL command 'CDF'
RETURN(t_stat, CDF(Tr_distn, t_stat), 1 - CDF(Tr_distn, t_stat));
```

This APPL code is implemented in a new APPL procedure called `CensoredT` and its use is illustrated in the example that follows. Assume there exists a medical treatment that has an established time-to-healing record that is modeled by the Gamma(2.1, 4.41) distribution, where time is measured in years. A new treatment is developed and experimenters hope to show an improvement (decrease) in healing time. The new treatment is administered to $n = 25$ patients, and it is noted that the first five healing times are 0.40, 0.54, 0.66, 0.75, 0.84 years. Completion of the full experiment, under the null hypothesis, has an expected time of $E(X_{(25)}) \doteq 4.52$ years, the expected healing time of the slowest patient to heal. However, the fifth patient's expected healing time, under the null hypothesis, is $E(X_{(5)}) \doteq 1.21$ years. Since the observed time of the fifth patient's healing was only 0.84 years, it would useful to know if there is enough statistical evidence to stop the experiment, concluding that the new treatment is better. The following APPL code will analyse this Type-II censored experiment:

```
> Old_Treatment := GammaRV(2.1, 4.41);
> n := 25;
> data := [0.40, 0.54, 0.66, 0.75, 0.84];
> CensoredT(Old_Treatment, data, n);
```

The procedure output is the test statistic, the lower tail p-value and the upper tail p-value. In this case those values are 1.309743, 0.031999, 0.968001. Since we are interested in the lower tail, we have a p-value of 0.031999, significant evidence that the new treatment is better and we can consider terminating the experiment.

4 Power simulation results

In this section, we discuss the results of various power simulations to see the effect of increasing r on the power of the test. We will use the T_n statistic and benchmark it against the A–D statistic. In the case of the Exponential and Normal prior distributions, power tests confirmed what was expected: as r increased, the power of the test increased, but never exceeded the power of letting all n components fail. In the case of the Gamma prior distribution, however, a non-intuitive result was observed. Power initially increased as r increased, but then started to decrease after reaching a 'maximum' power. Even more unexpectedly, in certain cases of parameter values, the maximum power of the mid-values of r was actually higher that the power of the full sample. We have investigated some of this unexpected phenomena and report on it below, as the implications of more power with lower r is very significant.

To implement this simulation, we wanted to set up the experiment so that, where possible, the underlying data had changing μ, but constant σ^2. We fixed σ^2 so that we could see if we could spot a change in μ by itself. This ability to detect a change is μ is helpful to lifetesters who have a new component that they would like to show superior to an existing component with a well defined distribution and well established μ. In the case of Exponentially distributed data, we could not fix σ^2 as $\sigma^2 = \mu^2$. We are able to fix σ^2 for the Normal and Gamma distributions. Table 1 shows the parameter values, as well as μ, and σ^2 for the Exponential, Normal, and Gamma distributions that were used in the power experiment.

Table 1: Distribution families, parameters, mean and variances for Monte Carlo Simulation

Normal Distribution, $H_0 : \mu = 1$, fixed $\sigma = 1$										
μ_a	-1	-0.8	-0.6	-0.4	-0.2	0.2	0.4	0.6	0.8	1
Exponential Distribution, $H_0 : \lambda = \mu = 1$										
$\lambda_a = \frac{1}{\mu_a}$	0.4	.6	0.7	0.8	0.9	1.25	1.5	1.9	2.3	2.7
Gamma Distribution, $H_0 : \alpha = \mu = 2.1$, $\beta = 4.41$ fixed $\sigma = 1$										
$\alpha_a = \mu_a$	1.1	1.3	1.5	1.7	1.9	2.3	2.5	2.7	2.9	3.1
β_a	1.21	1.69	2.25	2.89	3.61	5.29	6.25	7.29	8.41	9.61

As we see in the Normal and Exponential cases (Figures 1 and 2), higher r values produced higher power. Also, as can be seen in Figures 2 and 3, the T_r statistic produced higher power than the A–D statistic except for the extremely high values of μ (though not shown in Figure 1, the same result was observed for the Normal distribution). This switch is interesting since in the full sample experiments from the companion paper, The T_r appeared to always be higher in power than the A–D statistic.

A very counter-intuitive phenomena occurs with the Gamma distribution. Highest power for censored samples appears to come approximately $r = 10$ and then decrease as r approaches n. This result happened for the T_r and the A–D statistics. An enlargement of Figure 3 is shown in Figure 4 that further shows that the power increases then decreases. Figure 4 clearly shows that power starts out moderately at $r = 5$, then seems to achieve a maximum at $r = 10$ (for both statistics) then clearly decreases by the time $r = 20$ and $r = 25$. (Note the conditional order statistic approach at $r = n$ appears to be a different, less powerful statistic than the full sample for Gamma prior.) Most striking was that, for some values of μ_a lower than μ_0 we have achieved higher power for the censored, $r = 10$, case than we did for the full sample. As this is very counter-intuitive, we experimented in detail the case where the Gamma parameter $\alpha = 1.7$ and calculated the power for each value of $r = 1, 2, \ldots, 25$. The results of this in depth simulation are shown in Figure 5. Here we clearly see both phenomena occur: 1) power increases until approximate $r = 9$, then it decreases, and 2) for values for $r = 6, 7, 8, 9,$ and 10 power for the censored sample is at least has high or higher than power for the full sample. A note on the simulations: as these Gamma prior results were so counter-intuitive that our colleagues have had difficulty believing that a censored sample could possibly produce higher power than a full sample, we have re-designed and re-run this experiment a number of times over the last year, achieving similar results each time. For a copy of the simulation code, readers may contact the first author.

5 Applications and Implications

This methodology has potential for significant advances in reliability engineering lifetesting, pharmaceutical drug tests, or any sort of experiment where data comes naturally in ordered form. The sequential testing ability allows for a test to be terminated early, hence ending a dangerous experiment or giving early vindication allowing an effective therapy to go to market earlier. In particular, if a new therapy or component is more effective than the old, early failures may be remarkably small or large. This will result in acceptance and termination without running until all cases have failed. The test can then be used to accept the new component or medical treatment. Similarly, a few early failures can render a judgment and the remaining patients can be switched to potentially better therapies. Other implications of this research is as follows:

- Good statistical power for censored samples is possible for a wide ranges of experiments.

- Experiments can be designed for high n values, knowing that they will stop at a predetermined, relatively small r value.

- Experiments can be tracked real-time to see a pattern of p-values that indicates enough inference has been gained.

- With a Gamma prior distribution, higher power is achieved in censored samples than with full samples in some cases.

6 Conclusions

A new goodness of fit methodology has been developed and tested. Significant increases in power on the order of 30% have been found compared to the standard Anderson - Darling statistic. Also, relatively high power is achieved using the T_n statistic on censored samples, allowing for lifetests to be terminated early. Finally, in at lease one special case, that of a Gamma prior, a phenomena has been found, that at approximately $r = 0.4\,n$, power is greater that with a full sample.

7 Topics of Further Research

The cause of the phenomena revealed by validation testing of the slightly higher power in one special case needs to be further investigated. A possible basis for the explanation lies in the variance of successive, truncated order statistics, when data that originates from the alternate hypothesis is passed through the PIT of the Gamma distribution based on the null hypothesis. Somehow, the transformation of the data has a different characteristic than when is it passed thru a PIT based on a null hypothesis with, say, an Exponential prior. Also, further research is needed to investigate how high to set n and r in experimental design, in order to gain possible advantages in lower time on test, lower cost, and fewer failed items as a result of the experiment. For example, if a budget can afford 25 items failing, perhaps it would be more effective to put 50 items on test, knowing ahead of time that the desired increase on μ should be evident by about the $r = 10^{\text{th}}$ failure. Clearly a time savings and component savings is evident here. Finally, one of our goals was to find the exact power functions instead of using simulation of power. Due to the complexity of sending data from one distribution thru the PIT of another, the resulting transformations were so complicated that we could only find the exact power function for the Exponential prior with $r = 2$.

References

Balakrishnan, N., H. K. T. Ng, and N. Kannan, (2002), "A Test for Exponentiality," published in *Goodness-of-fit Tests and Model Validity*, editors C. Huber-Carol, N. Balakrishnan, M. S. Nikulin, and M. Mesbauh, Birkhaeuser.

David, H. A. (1981), *Order Statistics*, Second edition, John Wiley and Sons.

Glen, A., L. Leemis, and D. Barr , "Order Statistics in Goodness of Fit Testing," *IEEE Transactions on Reliability*, **50**, Number 2, 2001 pp. 209–213.

Glen, A., L. Leemis, and D. Evans , "APPL: A Probability Programming Language," *The American Statistician*, **55**, Number 2, 2001, pp. 156–166.

Glen, A., B. Foote , "Test for Uniformity based on Convolutions of the Uniform Distribution," *Technical Report, USMA* , West Point, NY, 2003.

Hegazy, Y., J. Green , "Some New Gooness-of-Fit Tests Using Order Statistics," *Applied Statistics*, **24**, Issue 3, 1975, pp. 299–308.

Maple Version 7, 2001, Waterloo Maple Inc., Waterloo, Canada. *Order Statistics*, Second edition, John Wiley and Sons.

Michael, J. R. and W. R. Schucany, "A New Approach to Testing Goodness of Fit for Censored Samples," *Technometrics*, **21**, Number 4, 1979.

O'Reilly, F. J. and M. A. Stephens, "Transforming Censored Samples for Testing Fit," *Technometrics*, **30**, Number 1, 1988.

Rosenblatt, M., "Remarks on a Multivariate Transformation," *Annals of Mathematical Statistics*, **23**, Issue 3, 1952.

Stephens, M. A., "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, **69**, Issue 347, 1974.

Figures

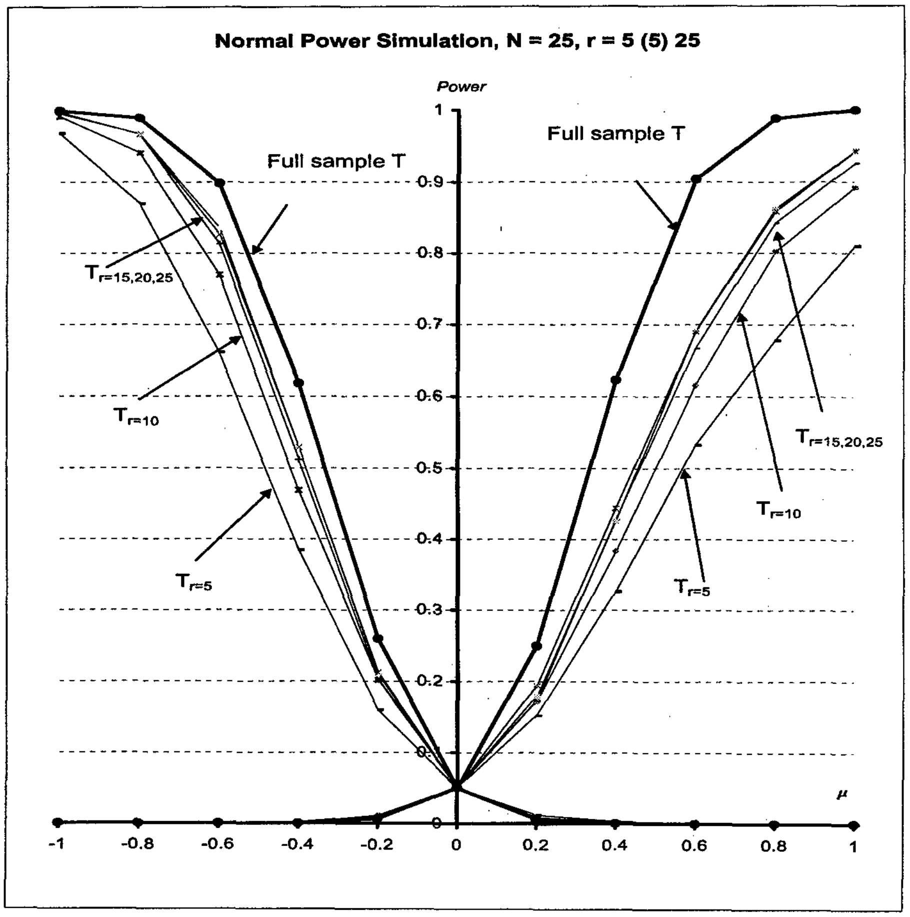

Figure 1: Results of Monte Carlo power simulation with underlying normally distributed data, $\sigma = 1$ and type I error $\alpha = 0.05$. Under H_0, $\mu = 0$. Only the T_n results are shown. Notice how well behaved the power functions are, in that higher r produced higher power.

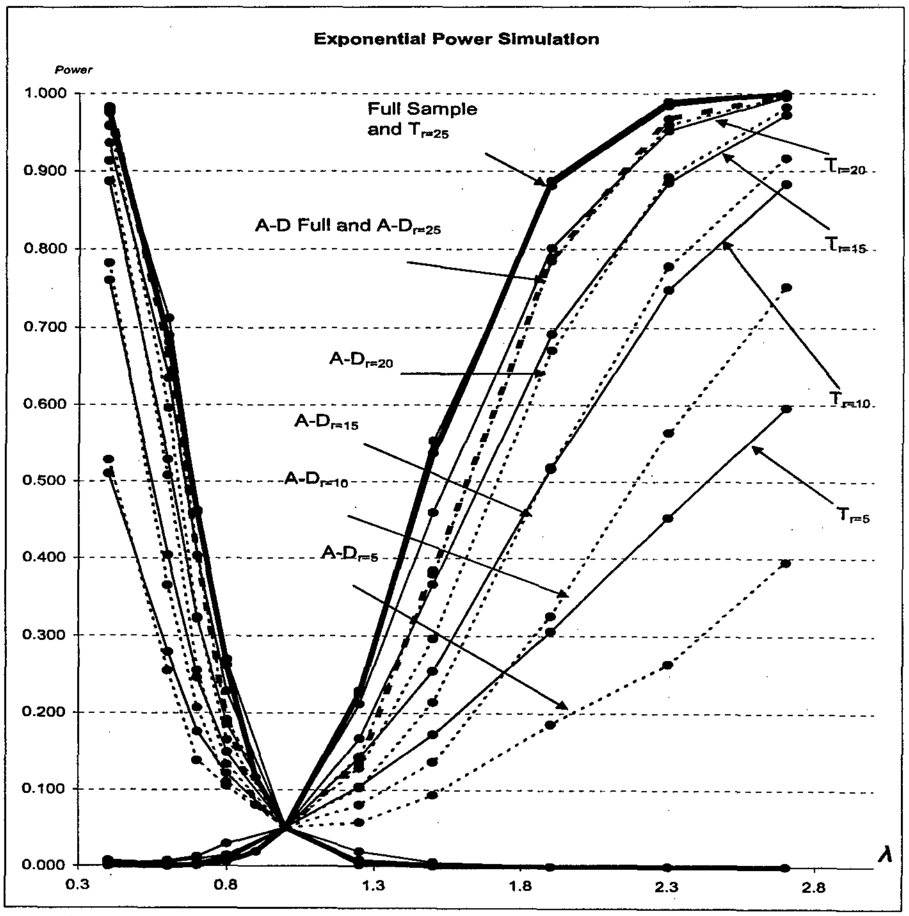

Figure 2: Results of Monte Carlo power simulation with underlying exponential distributed data with type I error $\alpha = 0.05$. Under H_0, the exponential distribution has parameter $\lambda = \frac{1}{\mu} = 1$. Thus, the upper tail test applies to the lower λ_a values. Notice that, like the Normal distribution, higher power is achieved for higher r values. Also notice how T_r achieves higher power than $A - D$, except for low values of λ_a (high values of μ_a).

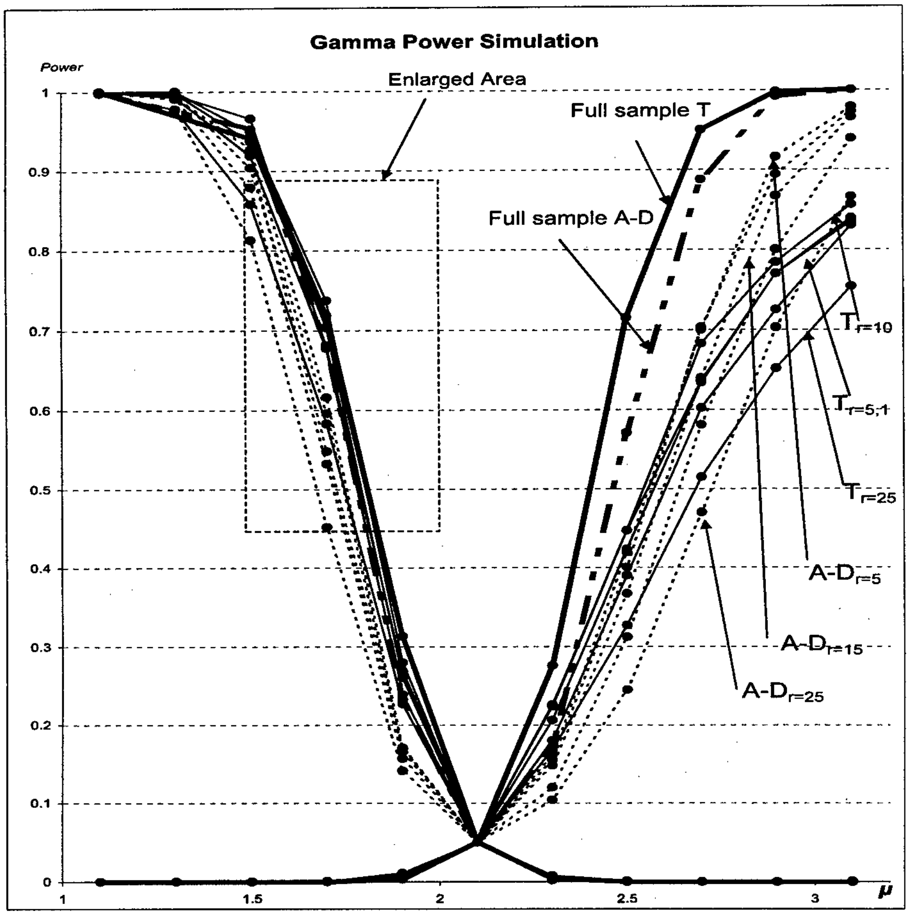

Figure 3: Results of Monte Carlo power simulation with underlying Gamma distributed data, $\sigma = 1$ and type I error $\alpha = 0.05$. Under H_0, the Gamma distribution has parameters $\alpha = \mu = 2.1$ and $\beta = 4.41$. Here the counter-intuitive result of higher power comes from $r = 10$ and then decreases as $r > 10$ for both the T_r and the $A - D$ test statistics. This phenomena is evident in the enlarged area shown in figure 4.

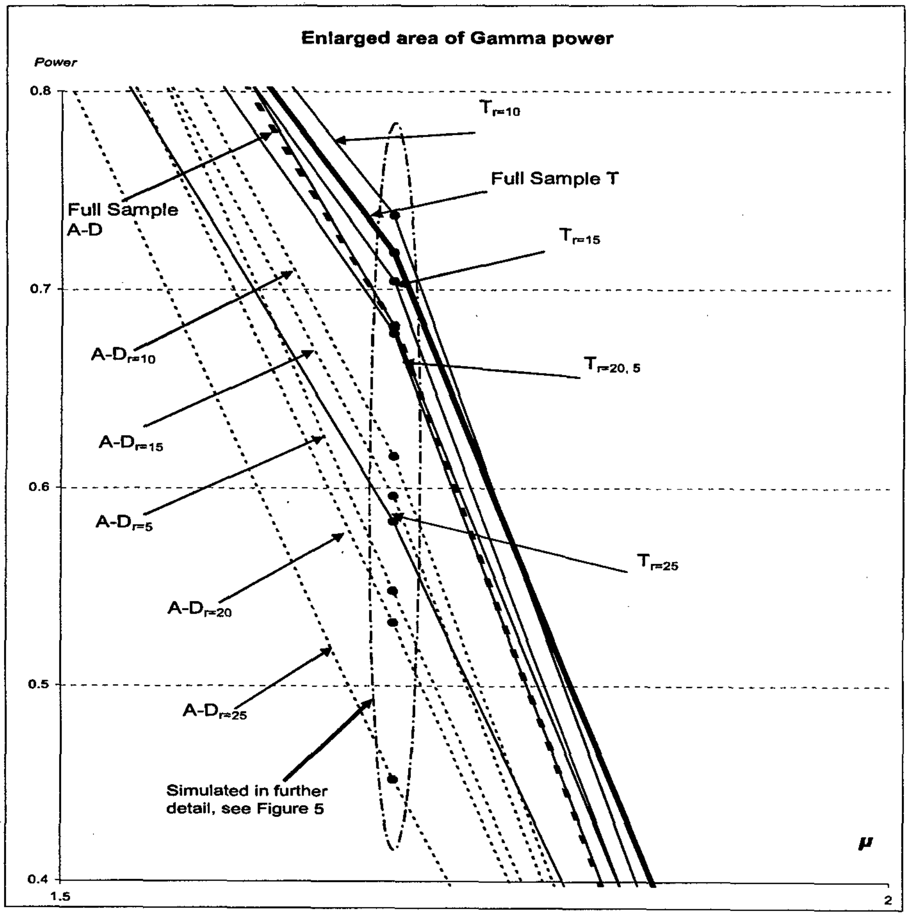

Figure 4: This enlarged area shows clearly the case that $T_{r=10}$ has higher power than even the full sample $T_{n=25}$. Thus we see the counter-intuitive result that under certain conditions an experiment can actually achieve higher power with a censored sample than with a full sample. Further investigation of this phenomenum at a higher resolution of r is found in Figure 5.

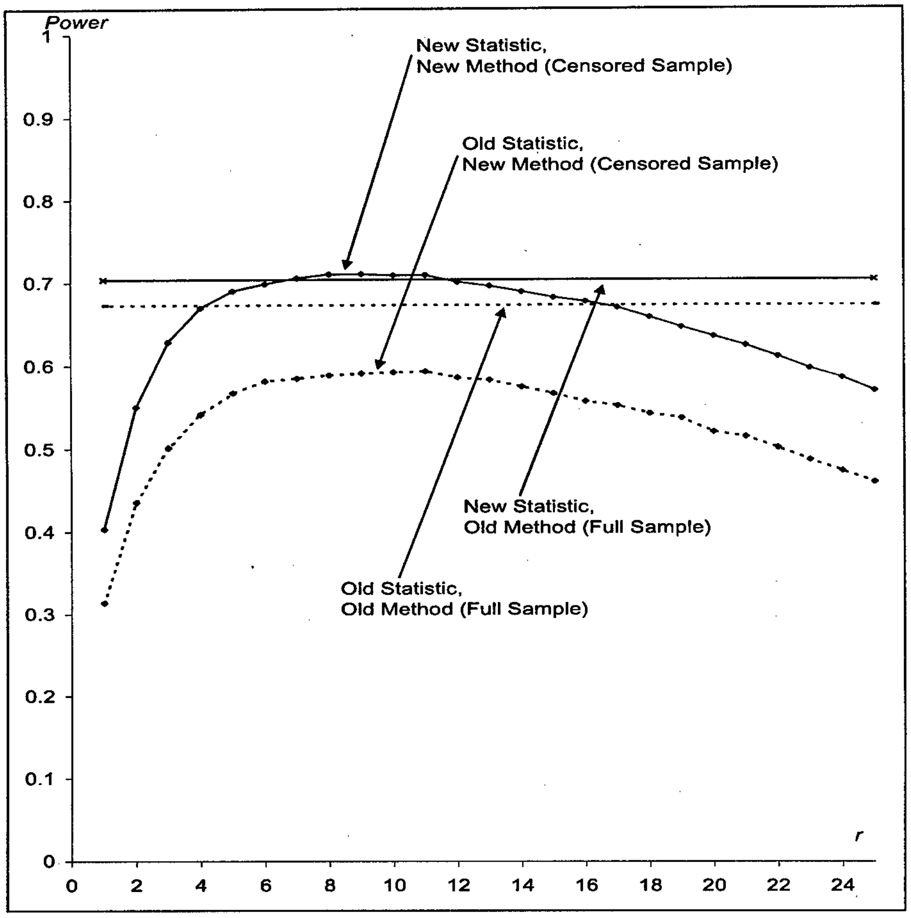

Figure 5: In an in depth experiment suggested from Figure 4, here is a plot of r versus power for each value of $r = 1$ (1) 25 for $\mu_a < \mu_0$. Note the two phenomena that 1) power increases on r then decreases for both statistics and 2) the special cases at $r = 6, 7, 8, 9,$ and 10 where higher power is achieved in a censored sample than with a full sample.

A Uniformity Test and an Inference Methodology for Goodness-of-Fit Lifetests with Type-II Right Censoring

Lt. Colonel Andrew G. Glen, Ph.D.	Bobbie Leon Foote, Ph.D.
Department of Mathematical Sciences	Department of Systems Engineering
United States Military Academy	United States Military Academy
West Point, New York 10996	West Point, New York 10996
aa1275@usma.edu	fb9690@usma.edu

September 26, 2003

ABSTRACT

We propose a methodology for gaining statistical inference during the actual conduct of the lifetest experiment that can reduce time on test and cost. We also present a test for uniformity based on the convolution of iid uniform random variables that complements the new methodology, producing design efficiencies in lifetime testing. The method is distribution free for any fully specified continuous distribution under the null hypothesis, and the distribution of the test statistic is calculated analytically, thus producing exact p-values. In certain non-intuitive cases, the methodology and statistic provide higher power for censored samples than for the complete samples. We achieve significant increases in power over the benchmark Anderson-Darling statistic.

Keywords: Computational Algebra Systems, Conditional Order Statistics.

1 Introduction and Literature Review

In lifetesting applications, tests are designed to gain an understanding of the probabilistic properties of a new drug, treatment, mechanical component or a system of components. Often, the costs of lifetests, in both time and money, constrain the design of the experiment, limiting the number of items placed on test and the length of the test. Many times, such as in the cases of pharmaceutical drug testing, the length of the experiment cannot be estimated accurately in advance, and often one is faced with un-analyzed, censored data in an ongoing experiment. For such cases we propose a methodology that gives exact statistical inference on censored samples. Consider an existing drug, process, or component with an all-parameters known lifetime reliability distribution $F(x)$. Should an improved drug, process, or component come along, both producers and consumers would like to verify that the new item is better than the existing item, most often by determining if its mean lifetime has improved (whether a decrease or an increase). In the lifetesting of the new item, it would be highly desirable to stop the test when enough evidence exists to support either claim. Such censoring, commonly called Type I (stop after time t) or Type II (stop after r items fail), can produce statistical inference, however, existing methods are not widely known, nor do they have remarkable statistical power. We propose a methodology that will specifically rely on Type II right censoring in the design and conduct of the lifetest. If for example, one could afford a lifetest with $n = 5$ items to fail, a certain level of statistical power could be achieved if the test continued until completion of n failures. Consider, however, an example where $n = 25$ items are placed on test with $r = 5$ as the designated censoring value. Obviously the second test would conclude more quickly, as the expected time on test would be the mean failure time of $X_{(25:5)}$, the fifth order statistic from a sample of 25 items, under the null hypothesis. Now consider a slightly different example, where $n = 25$ items are placed on test. Experimenters notice that after $r = 3$ failures, lifetimes seem to be substantially better than the original system. After $r = 6$ failures, they are convinced, at least anecdotally, that the new system is better. We propose a new methodology and a new test statistic that will allow for instantaneous assessment at every failure, with exact p-values, from an exact distribution of the test statistic. We rely on properties of conditional order statistic distributions to provide statistical inference for censored data that has acceptable statistical power. We also rely on the advances of computer algebra systems, especially A Probability Programming Language (APPL, Glen et. al., 2001) as our technique requires calculating many CDFs of these conditional order statistics as well as the very complicated distribution of the test statistic, all of which result in exact p-values for inference. We also show that for the case of the prior Gamma distribution, given certain conditions, it is possible to achieve higher power with a censored sample than it is for a full sample, a counter-intuitive result that has warranted in depth investigation on our part. The method we propose transforms either a full data set or a censored data set, via two probability integral transforms (PITs) and conditional order statistics, into an un-ordered, iid sample of uniformly distributed data on the open interval (0,1), which we abbreviate $U(0, 1)$. The test statistic T_n, based on the sum of $U(0, 1)$ random variables (rvs) and designed as a test of uniformity, enjoys significantly higher power than the A^2 statistic when finding differences in the mean of the distribution of the item in question, thus higher power is generally possible by combining the censored methodology with the use of the T_n statistic. The net effect of combining the new statistic with the new methodology is a very strong advantage in assessing censored data, that includes the possibility of purposefully designing lifetests with higher values of n

so that the test can be censored early at a reasonable value of r, saving time, money, and items that were destroyed during the test.

Testing for uniformity in a sample has many applications, many of which are explained in Chapter 8 of *Goodness-of-fit Techniques*, (D'Agostino and Stephens, 1988). Rosenblatt (1952) presents theory that transforms joint conditional rvs to ordered, uniformly distributed rvs for the censored case (we will instead transform censored data to a complete un-ordered set of uniform data). David (1981) discusses the Markov nature of conditional order statistics. He explains the ability to convert conditional order statistics into specific truncated order statistics, a result that we will use as part of our method. O'Reilly and Stephens (1988) use a Rosenblatt transform, then invert that transformed data to test ordered uniform data (we will test un-ordered uniform data). Hegazy and Green (1975) present work on goodness-of-fit using expected values of order statistics with approximations used for critical values. Balakrishnan, Ng, and Kannan (2002) present a test for exponentiality that is based on progressively censored data, which uses a T statistic, however this statistic and this method is unrelated to the T_n statistic and the sequentially censored data analysis that we use. Michael and Schucany (1979) also present a transformation that takes censored data and transforms it into ordered uniform data. Since Michael as well as Stephens (1974) also point out that the A^2 statistic is generally more powerful than the other well-known goodness-of-fit statistics in the case when the mean has shifted, we will compare the power of T_n with that of A^2 to show even higher power in detecting shifts in the mean than A^2.

2 The Test Statistic

The test statistic we propose, T_n, has the distribution of the convolution of iid $U(0,1)$ rvs:

$$T_n = \sum_{i=1}^{n} U_i = \sum_{i=1}^{n} F_X(X_i).$$

Prior to settling on this statistic, we explored other functions of iid $U(0,1)$ rvs. One option we explored was finding the distribution of $C = \sum_i^n \csc(U_i)$, as the cosecant function magnifies the statistic when the tails are too fat. The magnification happens at a quicker rate than that of $-\ln(U)$, and we found this statistic had slightly higher power than A^2, when testing for shifts in σ_a away from σ_0. The statistic had appreciably less power, though, when testing for changes in μ, a fact geometrically understandable, as the changes in μ do not exaggerate the test statistic quickly. Furthermore, the exact distribution of C could not be found, and critical points had to be estimated with Monte Carlo simulation, an inconvenience we wanted to avoid. We also considered $\min(U_i)$ and $\sum_{i=1}^n \tan(U_i)$ and found similarly unremarkable results.

We found considerable success with the test statistic $T_n = \sum_{i=1}^n U_i$ as a test for uniformity. Finding the distribution of the convolution of n iid $U(0,1)$ random variables becomes intractable by hand, once $n > 4$. However, by using APPL, we are able to determine the exact distribution of T_n for reasonable sample sizes. The distribution of T_n has n segments describing the PDF. Thus, the distribution of T_2 (the standard triangular distribution) has two segments, T_3 has three, and so on. As an example of a complete PDF not easily found by hand but possible in the APPL environment, the distribution of T_7 is as follows:

$$f(x) = \begin{cases} \frac{1}{720}\,x^6 & 0 < x < 1 \\ \frac{7}{120}\,x - \frac{7}{48}\,x^2 + \frac{7}{36}\,x^3 - \frac{7}{48}\,x^4 + \frac{7}{120}\,x^5 - \frac{1}{120}\,x^6 - \frac{7}{720} & 1 < x < 2 \\ \frac{1337}{720} - \frac{133}{24}\,x + \frac{329}{48}\,x^2 - \frac{161}{36}\,x^3 + \frac{77}{48}\,x^4 - \frac{7}{24}\,x^5 + \frac{1}{48}\,x^6 & 2 < x < 3 \\ -\frac{12089}{360} + \frac{196}{3}\,x - \frac{1253}{24}\,x^2 + \frac{196}{9}\,x^3 - \frac{119}{24}\,x^4 + \frac{7}{12}\,x^5 - \frac{1}{36}\,x^6 & 3 < x < 4 \\ \frac{59591}{360} - \frac{700}{3}\,x + \frac{3227}{24}\,x^2 - \frac{364}{9}\,x^3 + \frac{161}{24}\,x^4 - \frac{7}{12}\,x^5 + \frac{1}{48}\,x^6 & 4 < x < 5 \\ -\frac{208943}{720} + \frac{7525}{24}\,x - \frac{6671}{48}\,x^2 + \frac{1169}{36}\,x^3 - \frac{203}{48}\,x^4 + \frac{7}{24}\,x^5 - \frac{1}{120}\,x^6 & 5 < x < 6 \\ \frac{117649}{720} - \frac{16807}{120}\,x + \frac{2401}{48}\,x^2 - \frac{343}{36}\,x^3 + \frac{49}{48}\,x^4 - \frac{7}{120}\,x^5 + \frac{1}{720}\,x^6 & 6 < x < 7\,. \end{cases}$$

Interestingly, the PDF of T_{50} requires 91 pages of ASCII text to express. The distribution of

T_n for $n \leq 50$ can be found at the first author's web site, `www.dean.usma.edu/math/people/glen`. Since the distribution of T_n is exactly known, the exact critical values are calculable, and exact significance levels are attainable for any sample of data. Tables of critical values used for our power simulations are available from the first author and are left out of this paper for brevity. However, as APPL can find exact p-values for these distributions, tables such as these are becoming less necessary.

3 The Methodology

Let the lifetime of an existing system (also that of the null hypothesis) be distributed by the all-parameters known continuous rv X with CDF $F(x)$. Let n items be on lifetest and let the Type II right censoring value be r. Recall that in a lifetest, failure data arrives in ordered fashion. The ordered lifetime data $x_{(i)}$ have CDFs from their appropriate order statistic $F_{X_{(n:i)}}(x_{(n:i)})$, $i = 1, 2, \ldots, r$, (note $X_{(n:i)}$ is abbreviated $X_{(i)}$). In his work on order statistics, David (1981, pg. 21) explains two useful properties that we employ. First, order statistics form a Markov chain, in that for $r < s$,

$$f_{X_{(s)}|X_{(r)}=x_{(r)},X_{(r-1)}=x_{(r-1)},\ldots,X_{(1)}=x_{(1)}}(y) = f_{X_{(s)}|X_{(r)}=x_{(r)}}(y).$$

Second, finding the distribution of these order statistics is made simpler with truncation. Theorem 2.7 on the same page of David's text explains "For a random sample of n from a continuous parent conditional distribution of $X_{(s)}$, given $X_{(r)} = x$ $(s > r)$, is just the distribution of the $(s-r)^{\text{th}}$ order statistic in a sample of $n-r$ drawn from $f(y)/[1-F(y)]$ $(y \geq x)$, i.e., from the parent distribution truncated on the left at x." Thus for our purposes the CDF of the i^{th} order statistic, given the $(i-1)^{\text{th}}$ data point, $F(x_{(i)}|x_{(i-1)})$, is that of the rv $X_{(n-i+1:1)}$ with support $x_{(i-1)} < x_{(i)} < 1$. David shows this is the first order statistic from a sample size $n-(i-1)$ from the parent distribution of X truncated on the left at

$x_{(i-1)}$. In other words, the distribution is independent of $x_{(1)}, x_{(2)}, \ldots, x_{(i-3)}$, and $x_{(i-2)}$, and is therefore memoryless in this regard. Since each of the conditional distributions can be computed, conducting separate PITs on each data value, $F_{X_{(i)}|X_{(i-1)}}(x_{(i)})$, $i = 2, 3, \ldots, r$ will give a sample of r iid $U(0,1)$ random variables (see Rosenblatt 1952, pg. 470) to which a uniformity test can be applied. As mentioned earlier, we use T_n, as it is better at finding changes in μ_X than A^2 in many cases. So we will consider this statistic, renamed T_r to denote that it comes from a censored sample, which is shown to also have the distribution of the convolution of r iid $U(0,1)$ random variables. Therefore, T_r is found as follows:

$$T_r = \sum_{i=1}^{r} F_{X_{(i)}|X_{(i-1)}}(x_{(i)}) \, ,$$

where $F_{X_{(1)}|X_{(0)}}$ is defined to be $F_{X_{(1)}}$, and r is the size of the censored sample.

4 Implementation using APPL

The theory of the statistic is straightforward, however due to the requirement for multiple exact distribution functions that have support values which depend on the data, the implementation is made practicable only with automated probabilistic software. We implement the new method and new statistic in APPL (Glen, et. al. 2001) for a number of reasons. The software allows us to use exact distributions of the original data, the distributions of the conditional order statistics, and the distribution of the T_n statistic so that exact p-values are attainable. Additionally, the author has already calculated the PDFs of the sum of n $U(0,1)$ random variables from $n = 1$ to $n = 50$. APPL reads these PDFs exactly and can thus compute the exact p-values. APPL allows for the use of any continuous distribution (well-known distributions as well as ad hoc) to specify the null hypothesis and conducts the necessary PITs for these distributions. APPL calculates the CDFs of order statistics as well as truncated distributions. We will demonstrate the power of the censored and full samples using T_n and A^2 statistics with data from the Normal, Exponential, and Gamma prior distributions, however we are not limited to just these distributions.

The methodology can be confusing to those not used to using conditional order statistics, thus we present more clearly the algorithm for computing the test statistic.

- Specify the null distribution of the existing (old) system, $F(x)$.
- During the lifetest experiment, note n and create the vector of r observed occurrences.
- Calculate $z_{(i)} = F(x_{(i)}),\ i = 1, 2, \ldots, r$, which is ordered uniform (not iid).
- Calculate the unordered, iid $U(0, 1)$ (under the null hypothesis) $u_i = F_{Z_{(i)}|Z_{(i-1)}}(z_{(i)}),\ i = 1, 2, \ldots, r$. Note: we perform the PIT with $F(x)$ and then conduct the conditional order statistics PIT using the uniform conditional order statistic distributions. These two methods have been shown to be equivalent (Glen, et. al., 2001), but this method is preferred as the conditional order statistics of the uniform distributions are much more tractable than conditional order statistics using the parent distribution F. Also note, we find the conditional order statistic using the truncation of the parent distribution method outlined by David (1981).
- Sum the u_i values to get the T_r statistic.
- Calculate the p-value with the appropriate tail of the T_r distribution.

The APPL code that enacts this algorithm to calculate the statistic is as follows:

```
# take the r censored values in ‘data’ and PIT them into the list ‘Zdata’
for i from 1 to r do
  Zdata := [op(Zdata), CDF(Nulldist, data[i])];
od;
# sum the independent uniforms to for the statistic ‘t_stat’ starting with the first failure ...
```

```
t_stat:=CDF(OrderStat(U(0, 1), n, 1), Zdata[1]);
# ... then adding up the subsequent failures until r is reached.
if (r > 1) then
  for i from 2 to r do
    t_stat := t_stat + CDF(OrderStat(Truncate(U(0,1), evalf(Zdata[i-1]), 1),
      n - (i - 1), 1), Zdata[i]);
  end do; end if;
Tr_distn := cat('T',r);
# now return the statistic, the lower and upper tail p-values using APPL's 'CDF' command
RETURN(t_stat, CDF(Tr_distn, t_stat), 1 - CDF(Tr_distn, t_stat));
```

This algorithm is implemented in a new APPL procedure called CensoredT and its use is illustrated in the example that follows. Assume there exists a medical treatment that has an established time-to-recover record that is modeled by the Gamma(2.1, 4.41) distribution, where time is measured in years. A new treatment is developed and experimenters hope to show an improvement (decrease) in recover time. The new treatment is administered to $n = 25$ patients, and it is noted that the first five recover times are 0.40, 0.54, 0.66, 0.75, 0.84 years. Completion of the full experiment, under the null hypothesis, has an expected time of $E(X_{(25)}) \approx 4.52$ years, the expected recover time of the slowest patient to heal. However, the fifth patient's expected recover time, under the null hypothesis, is $E(X_{(5)}) \approx 1.21$ years. Since the observed time of the fifth patient's recover was only 0.84 years, it would useful to know if there is enough statistical evidence to stop the experiment, concluding that the new treatment is better. The following APPL code will analyze this Type-II censored experiment:

```
> Old_Treatment := GammaRV(2.1, 4.41);
> n := 25; data := [0.40, 0.54, 0.66, 0.75, 0.84];
> CensoredT(Old_Treatment, data, n);
```

The procedure output is the test statistic, the lower tail p-value and the upper tail p-value. In this case those values are 1.309743, 0.031999, 0.968001. Since we are interested in the lower tail, we have a p-value of 0.031999, significant evidence that the new treatment is better and we can consider terminating the experiment. For the exact calculations, to include the CDFs needed to compute this p-value, see the Appendix.

5 Power Simulation Results

In this section we discuss the results from extensive power experiments that compare full samples and censored samples of the T_n and T_r statistics versus the A^2 statistic. We rely on previous power studies by Stephens (1974) and Michael and Schucany (1979) that establish the A^2 statistic as generally more powerful than other statistics in testing for uniformity, especially when detecting a shift in μ. These other statistics include the Kolmogorov-Smirnov D, D^+, and D^- statistics, the Cramér-von Mises W^2 statistic, the Kupier V statistic, and the Watson U^2 statistic. As the A^2 statistic is generally more powerful than these statistics, we opine that it is sufficient to benchmark the new T_n statistic against the A^2 to show even more significant increases in power. We are interested in finding shifts in μ_0 given the all parameters known null hypothesis that X has CDF $F_X(x)$. As our intended purpose for the test statistic is to assist in the lifetesting task of finding changes in mean lifetime of items on test, this power simulation varies μ_a from μ_0, but fixes the standard deviation, where possible. We chose the Normal, Exponential and Gamma distributions as parent distributions for our null hypotheses. In the case of the Normal and Gamma distributions, we fixed $\sigma_0 = 1$ and only varied μ above and below μ_0. In the case of the Exponential, we have no choice but to vary μ_a and σ_a, as both depend on the parameter λ. We were interested in performing a similar simulation with the Weibull distribution, however, fixing σ and varying μ is not always possible, since the Weibull parameters are not readily solved for in such a manner, as not all combinations of mean and variance are possible with the Weibull distribution. Table 1 gives the various parameters, means, and variances, for each simulation of the Normal, Exponential, and Gamma null hypothesis distribution. Figures 1 thru 5 give the graphical results of our simulation. The T_n statistic for the full sample is denoted with the solid bold lines, while the T_r statistic for censored samples is denoted with the solid non-bolded lines. The A^2 full sample is denoted by the dashed bolded lines, while the A^2 censored sample is denoted by the dashed non-bolded lines. For the full sample simulation we chose sample sizes of $n = 5$ and $n = 25$ and compared the results of the A^2 and T_n tests. The results show significant improvement in power for all three distributions. For example, in the Normal power experiment, where $n = 5$ and $\mu_a = 0.6$, the power for A^2 is 26% and the power for T_n is 37%, a 42% increase in power. Similarly in the Gamma experiment, where $n = 25$ and $\mu_a = 2.3$, the power increase is from 17% to 28%, a 64% increase in power over the A^2 statistic. The full sample cases where $n = 25$ are shown in bold lines in Figures 1 thru 5, however the cases where $n = 5$ were omitted form the figures to avoid clutter. All tabular results from the simulations are available from the first author.

As we see in the Normal and Exponential cases (Figures 1 and 2), higher r values produced higher power. Also, as can be seen in Figures 2 and 3, the T_r statistic produced higher power than the A^2 statistic except for the extremely high values of μ (though not shown in Figure 1, the same result was observed for the Normal distribution). This switch is interesting since in the full sample experiments from the companion paper, The T_r appeared to always be higher in power than the A^2 statistic.

A very counter-intuitive phenomena occurs with the prior Gamma distribution. The highest power for censored samples appears to come at approximately $r = 10$ and then decreases as r approaches n. This result happened for the T_r and the A^2 statistics. An enlargement of Figure 3 is shown in Figure 4 that further shows that the power increases then decreases. Figure 4 clearly shows that power starts out moderately at $r = 5$, then seems to achieve a maximum at $r = 10$ (for both statistics) then clearly decreases by the time $r = 20$ and $r = 25$. (Note the conditional order statistic approach at $r = n$ appears to be a different, less powerful statistic than the full sample for Gamma prior.) Most striking was that, for some values of μ_a lower than μ_0 we have achieved higher power for the censored, $r = 10$, case than we did for the full sample. As this is very counter-intuitive, we experimented in detail for the case where the Gamma parameter was $\alpha = 1.7$ and calculated the power for each value of $r = 1, 2, \ldots, 25$. The results of this in depth simulation are shown in Figure 5. Here we clearly see both phenomena occur: 1) power increases until approximately $r = 9$, then it decreases, and 2) for $6 \leq r \leq 10$ power for the censored sample is at least has high or higher than power for the full sample. A note on the simulations: as these Gamma prior results were so counter-intuitive, we have re-designed and re-run this experiment a number of times over the last year, achieving similar results each time. For a copy of the simulation code, readers may contact the first author.

An important note involves the complexity of this Monte Carlo simulation. The simulation requires, among other things, the CDF of the prior distributions (non-trivial in the case of the Gamma), the ability to conduct a PIT on the data, the exact CDFs of truncated order statistics created from the data, and the exact distribution of the test statistic. None of this is currently possible outside of a computer algebra system, certainly not in well known statistics software packages. To see some of these distributions and to get an appreciation for the simplest case of $r = 5$ see the example in the Appendix.

6 Applications and Implications

This methodology has potential for significant advances in medical and reliability engineering lifetesting, pharmaceutical drug tests, or any sort of experiment where data comes naturally in ordered form. The sequential testing ability allows for a test to be terminated early, hence ending a dangerous experiment or giving early vindication allowing an effective therapy to go to market earlier. In particular, if a new therapy or component is more effective than the old, early failures may be remarkably small or large. This will result in acceptance and termination without running until all cases have failed. The test can then be used to accept the new component or medical treatment. Similarly, a few early failures can render a judgment and the remaining patients can be switched to potentially better therapies. Other implications of this research are as follows:

- Good statistical power for censored samples is possible for a wide range of experiments.
- Experiments can be designed with intentionally large values of n, knowing that they will stop at a predetermined, relatively small value of r.
- Experiments can be tracked real-time to see a pattern of p-values that indicates enough inference has been gained.
- With a Gamma prior distribution, higher power is achieved in censored samples than with full samples in some cases.
- Computer algebra systems are adept enough for computing the multiple CDFs needed for such a statistic and methodology to be practical in its implementation.

7 Conclusions and Further Research

A new goodness of fit methodology and a new uniformity test statistic has been developed and tested. Significant increases in power have been found compared to the benchmark Anderson–Darling statistic. Additionally, exact p-values for this test statistic are achievable. Also, relatively high power is achieved using the T_n statistic on censored samples, allowing for lifetests to be terminated early. Finally, in at least one special case, that of a Gamma prior, a phenomena has been found, that at approximately $r = 0.4\,n$, power is greater than with a full sample. The cause of the phenomena in which censored samples give higher power than full samples needs to be further investigated. A possible basis for the explanation lies in the variance of successive, truncated order statistics, when data that originates from the alternate hypothesis is passed through the PIT of the prior Gamma distribution. Somehow, the transformation of the data has a different characteristic than when is it passed thru a PIT based on a null hypothesis with, say, a prior Exponential distribution. Also, further research is needed to investigate how high to set n and r in experimental design, in order to gain possible advantages in lower time on test, lower cost, and fewer failed items as a result of the experiment. For example, if a budget can afford 25 items failing, perhaps it would be more effective to put 50 items on test, knowing ahead of time that if the desired increase in μ is present, it should be evident by about the, say $r = 10^{\text{th}}$, failure and the test can be terminated around that point. Thus, the potential for time savings and component savings are evident. Finally, one of our goals was to find the exact power functions instead of using simulation to compute power. Due to the complexity of sending data from one distribution thru the PIT of another, the resulting transformations were so complicated that we could only find the exact power function for the Exponential prior with $r = 2$. Clearly there is a need for this new statistic and this new methodology as well as a need for further research in this area.

Appendix

In this section we will show the computations needed to find the p-value listed in the example of the "Implementations" section of the paper. Recall an existing treatment is known to have a recovery time X distributed according to the Gamma(2.1, 4.41) rv with PDF $f(x) = 2.565595\, x^{3.41} e^{-2.1\,x}$, $0 < x < \infty$, and CDF, calculated in APPL of,

$$\begin{aligned} F(x) \;\; = \;\; & \tfrac{100000}{66987458901}\, x^{\frac{41}{200}} 21^{\frac{41}{200}} 10^{\frac{159}{200}} e^{-\frac{21}{20}\,x} [82181\, \texttt{WhittakerM}(\tfrac{41}{200}, \tfrac{141}{200}, \tfrac{21}{10}\,x) \\ & +143220\, \texttt{WhittakerM}(\tfrac{41}{200}, \tfrac{141}{200}, \tfrac{21}{10}\,x)x + 132300\, \texttt{WhittakerM}(\tfrac{41}{200}, \tfrac{141}{200}, \tfrac{21}{10}\,x)x^2 \\ & -71610\, x\, \texttt{WhittakerM}(-\tfrac{159}{200}, \tfrac{141}{200}, \tfrac{21}{10}\,x) - 82181\, \texttt{WhittakerM}(-\tfrac{159}{200}, \tfrac{141}{200}, \tfrac{21}{10}\,x) \\ & +92610\, x^3\, \texttt{WhittakerM}(\tfrac{41}{200}, \tfrac{141}{200}, \tfrac{21}{10}\,x) - 44100\, x^2\, \texttt{WhittakerM}(-\tfrac{159}{200}, \tfrac{141}{200}, \tfrac{21}{10}\,x)]/\Gamma(\tfrac{41}{100}), \end{aligned}$$

$0 < x$, relying on Maple's `WhittakerM` function, a solution to a differential equation. An experimental treatment is given to $n = 25$ patients in hopes of showing an improved recovery time. The null hypothesis is that $F_{new} = F_{old}$ and the alternate hypothesis is that $F_{new} < F_{old}$. The first five recovery times (in years) are noted to be 0.40, 0.54, 0.66, 0.75, 0.84 years. The algorithm first transforms these data values thru the Gamma CDF, $z_{(i)} = F(x_{(i)})$, $i = 1, 2, \ldots, 5$ (the first PIT) to come up with five transformed values 0.005202791275, 0.01549350642, 0.03082310670, 0.04677731189, 0.06666297196.

For each of the $z_{(i)}$ it is necessary to calculate the appropriate conditional order statistic CDF so that the unordered uniform variates can be calculated. The first data point has CDF

$$\begin{aligned} F_{X_{(1)}}(x) \;\; = \;\; & x^{25} - 25\, x^{24} + 300\, x^{23} - 2300\, x^{22} + 12650\, x^{21} - 53130\, x^{20} + 177100\, x^{19} \\ & -480700\, x^{18} + 1081575 x^{17} - 2042975\, x^{16} + 3268760\, x^{15} - 4457400\, x^{14} \\ & +5200300\, x^{13} - 5200300\, x^{12} + 4457400\, x^{11} - 3268760\, x^{10} + 2042975\, x^{9} \\ & -1081575\, x^{8} + 480700\, x^{7} - 177100\, x^{6} + 53130\, x^{5} - 12650\, x^{4} + 2300\, x^{3} - 300\, x^{2} \\ & +25\, x, \end{aligned}$$

for $0 < x < 1$. Thus the first iid U(0,1) p-value (from the second PIT) is $p1 = F_{X_{(1)}}(0.005202791275) = 0.1222639202$. The second data point comes from the truncated, order statistic from the uniform(0, 1) distribution with $n = 19$ and $r = 1$ truncated on the left at $z_{(1)} = 0.005202791275$. The CDF for this point is

$$\begin{aligned} F(x) &= -1.1334\,x^{24} + 27.2008\,x^{23} - 312.8093\,x^{22} + 2293.9350\,x^{21} - 12043.1589\,x^{20} \\ &\quad +48172.63\,x^{19} - 152546.68\,x^{18} + 392262.89\,x^{17} - 833558.64\,x^{16} + 1481882.03\,x^{15} \\ &\quad -2222823.1\,x^{14} + 2829047.5\,x^{13} - 3064801.48\,x^{12} + 2829047.52\,x^{11} - 2222823.05\,x^{10} \\ &\quad +1481882.03\,x^{9} - 833558.64\,x^{8} + 392262.89\,x^{7} - 152546.68\,x^{6} + 48172.64\,x^{5} \\ &\quad -12043.158\,x^{4} + 2293.935\,x^{3} - 312.809\,x^{2} + 27.201\,x - 0.1334, \end{aligned}$$

for $0.005202791275 < x < 1$. The second p-value is $p2 = F(0.01549350642) = 0.2208579439$. Similarly the third data point comes from the truncated, order statistic uniform(0, 1) distribution from $n = 18$ and $r = 1$ truncated on the left at $z_{(2)} = 0.01549350642$. The CDF for this point is

$$\begin{aligned} F(x) &= 1.4320\,x^{23} - 32.9382\,x^{22} + 362.3206\,x^{21} - 2536.2444\,x^{20} + 12681.2221\,x^{19} \\ &\quad -48188.644\,x^{18} + 144565.93\,x^{17} - 351088.693\,x^{16} + 702177.387\,x^{15} - 1170295.64\,x^{14} \\ &\quad +1638413.90\,x^{13} - 1936307.34\,x^{12} + 1936307.34\,x^{11} - 1638413.90\,x^{10} + 1170295.64\,x^{9} \\ &\quad -702177.387\,x^{8} + 351088.693\,x^{7} - 144565.933\,x^{6} + 48188.64\,x^{5} - 12681.22\,x^{4} \\ &\quad +2536.2444\,x^{3} - 362.3206\,x^{2} + 32.9382\,x - 0.4321, \end{aligned}$$

for $0.01549350642 < x < 1$. The third p-value is $p3 = F(0.03082310670) = 0.3029840349$. Likewise, the fourth data point comes from CDF

$$\begin{aligned} F(x) &= -1.9913\,x^{22} + 43.8082\,x^{21} - 459.9860\,x^{20} + 3066.5750\,x^{19} - 14566.2312\,x^{18} \\ &\quad +52438.432\,x^{17} - 148575.559\,x^{16} + 339601.278\,x^{15} - 636752.396\,x^{14} + 990503.727\,x^{13} \\ &\quad -1287654.85\,x^{12} + 1404714.37\,x^{11} - 1287654.85\,x^{10} + 990503.73\,x^{9} - 636752.396\,x^{8} \\ &\quad +339601.2779\,x^{7} - 148575.5591\,x^{6} + 52438.4326\,x^{5} - 14566.2313\,x^{4} + 3066.5750\,x^{3} \\ &\quad -459.98625\,x^{2} + 43.8082\,x - 0.99128, \end{aligned}$$

for $0.03082310670 < x < 1$. The fourth p-value is $p4 = F(0.04677731189) = 0.305920807$.

Finally, the fifth data point comes from the CDF

$$\begin{aligned} F(x) \quad = \quad & 2.7348\,x^{21} - 57.4298\,x^{20} + 574.2983\,x^{19} - 3637.2224\,x^{18} + 16367.500\,x^{17} \\ & -55649.502\,x^{16} + 148398.674\,x^{15} - 317997.158\,x^{14} + 556495.027\,x^{13} - 803826.151\,x^{12} \\ & +964591.381\,x^{11} - 964591.381\,x^{10} + 803826.151\,x^{9} - 556495.027\,x^{8} + 317997.158\,x^{7} \\ & -148398.674\,x^{6} + 55649.503\,x^{5} - 16367.5001\,x^{4} + 3637.222\,x^{3} - 574.2982\,x^{2} \\ & +57.4298\,x - 1.7347, \end{aligned}$$

for $0.04677731189 < x < 1$. The fifth p-value is $p5 = F(0.06666297196) = 0.357716701$. The test statistic is the sum of the iid uniforms, $t = p1 + p2 + p3 + p4 + p5 = 1.309743407$, and has a CDF based on the null hypothesis of that of convolution of five Uniform(0, 1) random variables:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{120}\,x^5 & 0 < x < 1 \\ \frac{1}{24} + \frac{5}{12}\,x^2 - \frac{5}{12}\,x^3 + \frac{5}{24}\,x^4 - \frac{1}{30}\,x^5 - \frac{5}{24}\,x & 1 < x < 2 \\ -\frac{21}{8} + \frac{155}{24}\,x - \frac{25}{4}\,x^2 + \frac{35}{12}\,x^3 - \frac{5}{8}\,x^4 + \frac{1}{20}\,x^5 & 2 < x < 3 \\ \frac{141}{8} - \frac{655}{24}\,x + \frac{65}{4}\,x^2 - \frac{55}{12}\,x^3 + \frac{5}{8}\,x^4 - \frac{1}{30}\,x^5 & 3 < x < 4 \\ -\frac{601}{24} + \frac{625}{24}\,x - \frac{125}{12}\,x^2 + \frac{25}{12}\,x^3 - \frac{5}{24}\,x^4 + \frac{1}{120}\,x^5 & 4 < x < 5 \\ 1 & 5 < x < \infty \end{cases}.$$

The lower tail p-value is therefore $F(1.309743407) = 0.0319993056$, a low p-value suggesting the recovery time for the new treatment is lower than that of the existing treatment.

References

Balakrishnan, N., Ng, H. K. T., and Kannan, N. (2002), "A Test for Exponentiality," published in *Goodness-of-fit Tests and Model Validity*, editors C. Huber-Carol, N. Balakrishnan, M. S. Nikulin, and M. Mesbauh, Birkhaeuser.

David, H. A. (1981), *Order Statistics*, Second edition, John Wiley and Sons.

Glen, A., Leemis, L., and Barr, D. (2001) "Order Statistics in Goodness of Fit Testing," *IEEE Transactions on Reliability*, **50**, Number 2, pp. 209–213.

Glen, A., Leemis, L., and Evans, D. (2001), "APPL: A Probability Programming Language," *The American Statistician*, **55**, Number 2, pp. 156–166.

Hegazy, Y., Green, J. (1975), "Some New Goodness-of-Fit Tests Using Order Statistics," *Applied Statistics*, **24**, Issue 3, pp. 299–308.

Maple Version 7 (2001), Waterloo Maple Inc., Waterloo, Canada. *Order Statistics*, Second edition, John Wiley and Sons.

Michael, J. R. and Schucany, W. R. (1979), "A New Approach to Testing Goodness of Fit for Censored Samples," *Technometrics*, **21**, Number 4.

O'Reilly, F. J. and Stephens, M. A. (1988), "Transforming Censored Samples for Testing Fit," *Technometrics*, **30**, Number 1.

Rosenblatt, M. (1952), "Remarks on a Multivariate Transformation," *Annals of Mathematical Statistics*, **23**, Issue 3.

Stephens, M. A. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, **69**, Issue 347.

Tables and Figures

Table 1: Distribution families, parameters, mean and variances for Monte Carlo Simulation

Normal Distribution, $H_0 : \mu = 1$, fixed $\sigma = 1$										
μ_a	-1	-0.8	-0.6	-0.4	-0.2	0.2	0.4	0.6	0.8	1
Exponential Distribution, $H_0 : \lambda = \mu = 1$										
$\lambda_a = \frac{1}{\mu_a}$	0.4	.6	0.7	0.8	0.9	1.25	1.5	1.9	2.3	2.7
Gamma Distribution, $H_0 : \alpha = \mu = 2.1$, $\beta = 4.41$ fixed $\sigma = 1$										
$\alpha_a = \mu_a$	1.1	1.3	1.5	1.7	1.9	2.3	2.5	2.7	2.9	3.1
β_a	1.21	1.69	2.25	2.89	3.61	5.29	6.25	7.29	8.41	9.61

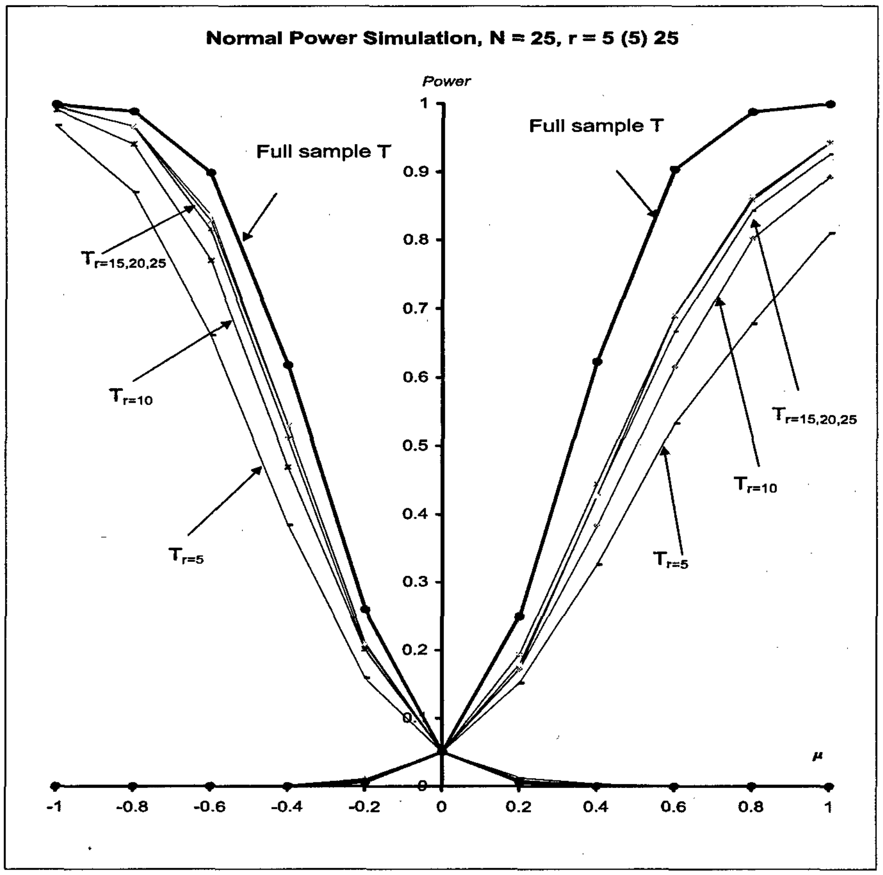

Figure 1: Results of Monte Carlo power simulation with underlying normally distributed data, $\sigma = 1$ and type I error $\alpha = 0.05$. Under H_0, $\mu = 0$. Only the T_n results are shown. Notice how well behaved the power functions are, in that higher r produced higher power.

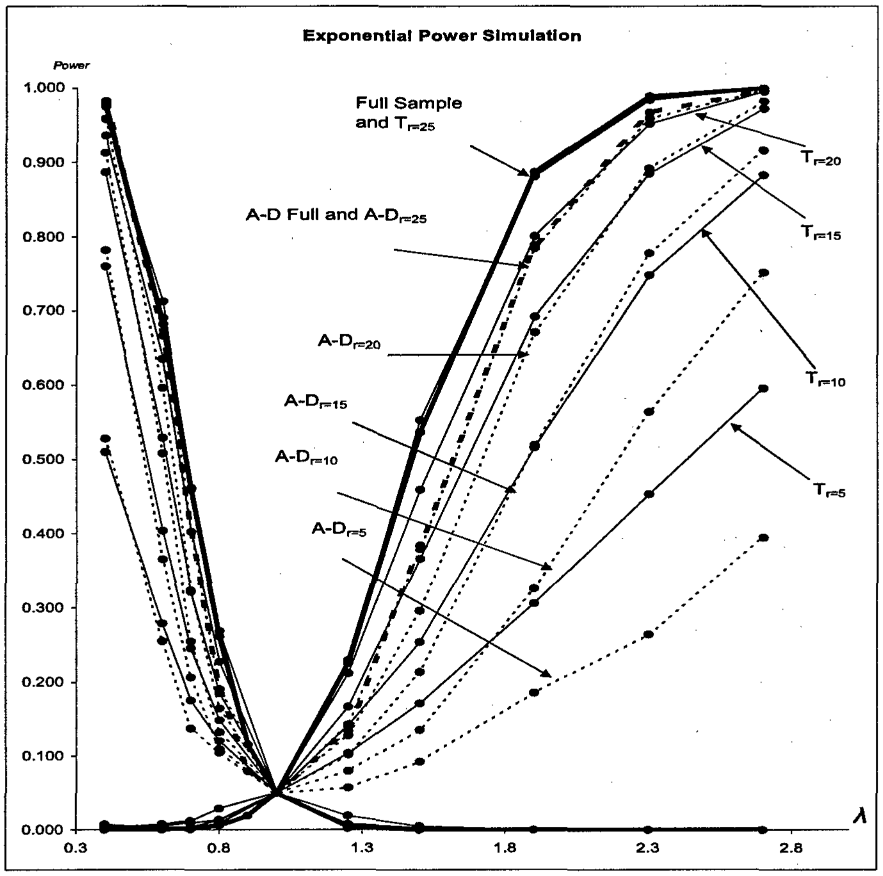

Figure 2: Results of Monte Carlo power simulation with underlying exponential distributed data with type I error $\alpha = 0.05$. Under H_0, the exponential distribution has parameter $\lambda = \frac{1}{\mu} = 1$. Thus, the upper tail test applies to the lower λ_a values. Notice that, like the Normal distribution, higher power is achieved for higher r values. Also notice how T_r achieves higher power than A^2, except for low values of λ_a (high values of μ_a).

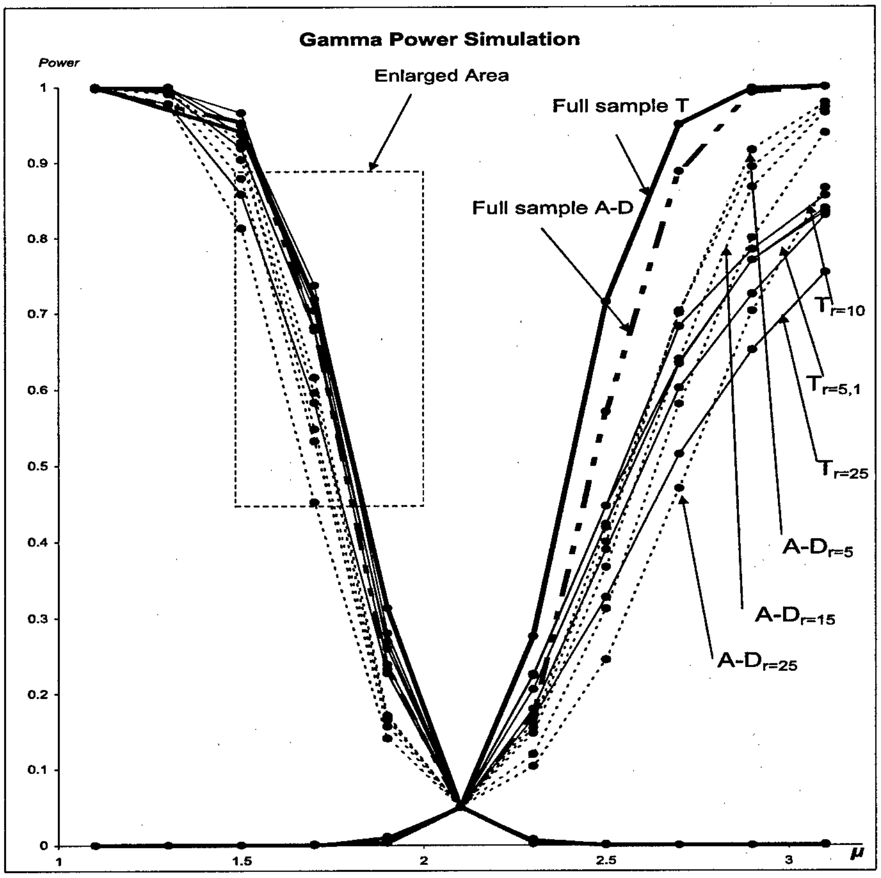

Figure 3: Results of Monte Carlo power simulation with underlying Gamma distributed data, $\sigma = 1$ and type I error $\alpha = 0.05$. Under H_0, the Gamma distribution has parameters $\alpha = \mu = 2.1$ and $\beta = 4.41$. Here the counter-intuitive result of higher power comes from $r = 10$ and then decreases as $r > 10$ for both the T_r and the A^2 test statistics. This phenomena is evident in the enlarged area shown in figure 4.

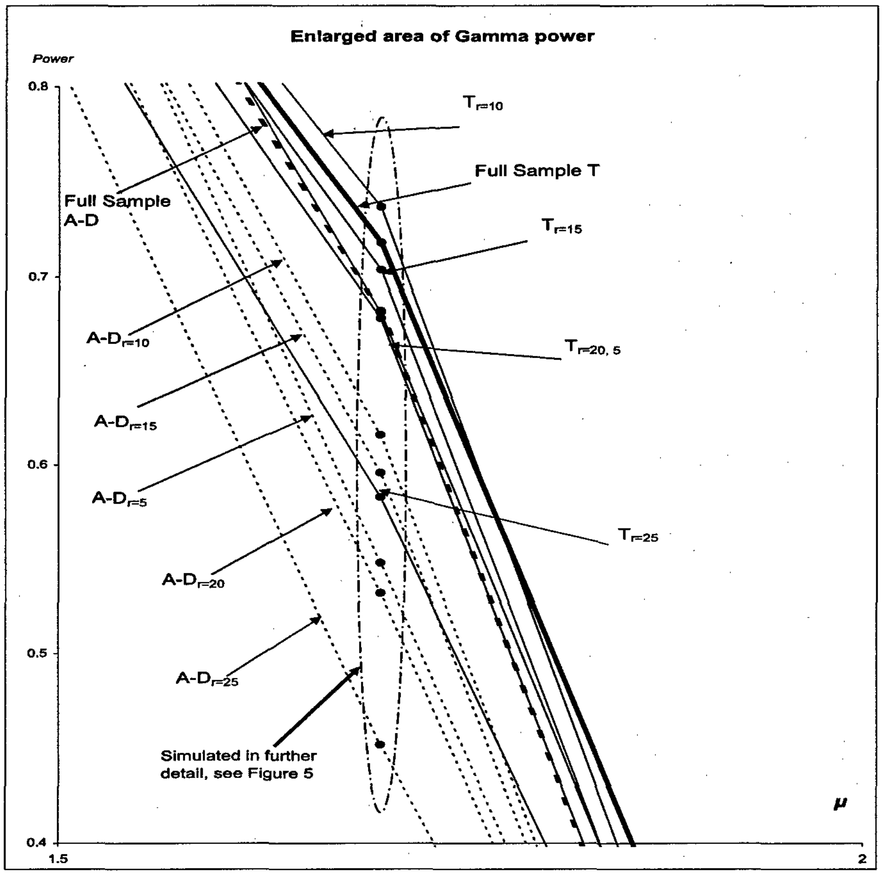

Figure 4: This enlarged area shows clearly the case that $T_{r=10}$ has higher power than even the full sample $T_{n=25}$. Thus we see the counter-intuitive result that under certain conditions an experiment can actually achieve higher power with a censored sample than with a full sample. Further investigation of this phenomenom at a higher resolution of r is found in Figure 5.

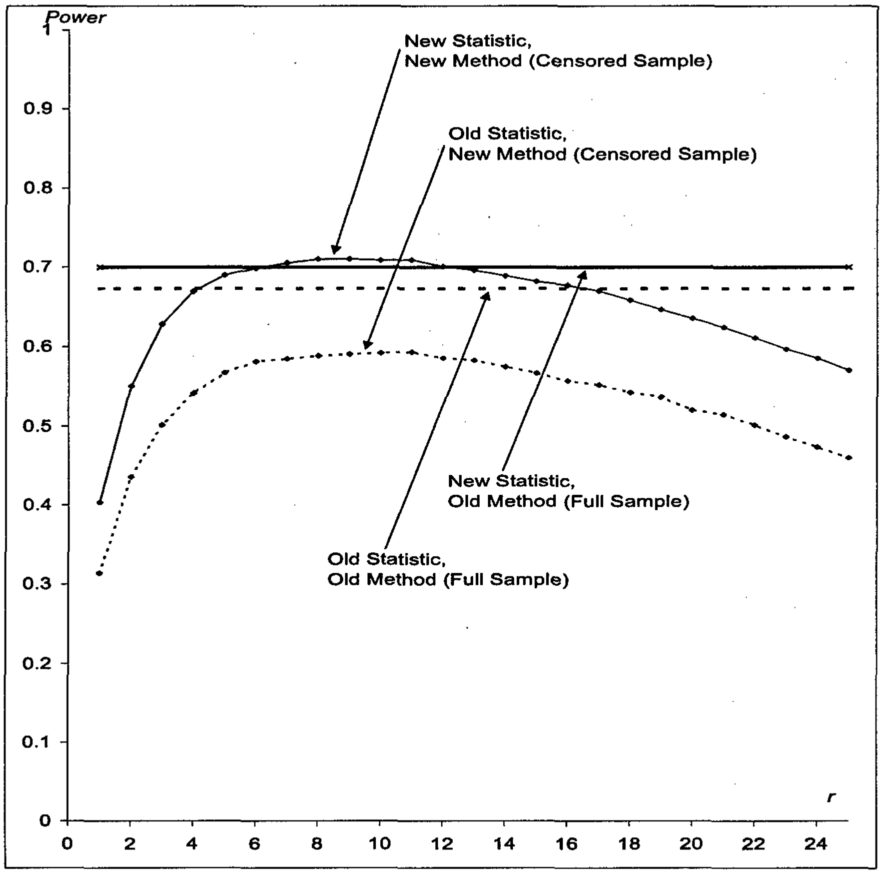

Figure 5: In an in depth experiment suggested from Figure 4, here is a plot of r versus power for each value of $r = 1\ (1)\ 25$ for $\mu_a < \mu_0$. Note the two phenomena that 1) power increases on r then decreases for both statistics and 2) the special cases at $r = 6, 7, 8, 9,$ and 10 where higher power is achieved in a censored sample than with a full sample.

The process for getting exact inference on right censored samples during a life test.

1. **Requirements**
   - Maple Software, Commercially available from Waterloo, Maple Inc. , Waterloo, Canada.
   - APPL Software, public domain software, available at www.usma.edu/math/people/glen
   - A working knowledge of basic probability, statistics, and reliability engineering.
   - A beginner's understanding of the use of Maple and APPL
2. **Process execution:**
   - Before the experiment, the experimenter has 1) a well defined , currently existing item/ process/ drug/ system with fully specified lifetime distribution function, and 2) a new item/ process/ drug/ system that is hopefully better than the current one.
   - The experimenter places n of the new items on life test, and records the failure times of the new items as they occur.
   - At each occurrence of failure, the experimenter enters the newly noted failure time into a list of previously noted failure times and re-executes the CensoredT command to get the most up-to-date statistical P-value so far .
   - When the statistical P-value is sufficiently small (often times smaller than 0.01) there is exact statistical evidence that the new item/process /drug/system is better than the current one.
3. **Example:**
   - A current light bulb has a well defined mean life of 1000 hours and is adequately described by the exponential random variable with parameter 1/1000. An experimenter wants to show that a new light bulb has a higher mean life.
   - The experimenter places n=35 new versions of light bulbs on life test and notes the following failure times. The first bulb fails at time 49 hours, the second at time 72 hours, the third at time 115 hours, and the fourth at time 197 hours.
   - The P values after each successive failure are in order: .255, .203, .113, .034.
   - At this point we have significant evidence(at least at the .05 level of significance) to conclude that the new light bulbs have a longer mean life.
   - We can thus stop the test at time 197 hours, as opposed to waiting for the total experiment to end, which would be at about 4146 hours, if the new light bulb is no better that the old(longer if the ones are better).
   - In a Maple work sheet, the steps to this example and the results are found on the following page.

```
> restart;

> read(`d:/APPL/appl.txt`); read(`d:\CensoredT.txt`);

> X:=ExponentialRV(1/1000.);

        X := [[x -> .001000000000e^(-.001000000000x)], [0, ∞], ["Continuous", "PDF"]]

> data:=[49];

                              data := [49]

> CensoredT(X,data,35);

                   .7446193253, .7446193253, .2553806747

> data:=[49,72];CensoredT(X, data,35);

                            data := [49, 72]

                   1.362546169, .7968263065, .2031736935

> data:=[49,72,115];CensoredT(X, data,35);

                          data := [49, 72, 115]

                    2.120590359, .886649768, .113350232

> data:=[49,72,115,197];CensoredT(X, data,35);

                        data := [49, 72, 115, 197]

                    3.048079409, .965786950, .034213050
```

The original distribution assumptions can be other than the exponential. However for Normal, Weibull, Gamma, and Beta distributions in particular cases the computational burden may be too high or has a result that cannot be evaluated. This process does not work for prior distribution assumptions that are discrete.

Distribution List

The list indicates the complete mailing address of the individuals and organizations receiving copies of the report and the number of copies received. Due to the Privacy Act, only use business addresses; no personal home addresses. Distribution lists provide a permanent record of initial distribution. The distribution information will include the following entries:

NAME/AGENCY	ADDRESS	COPIES
Author(s)	Department of Systems Engineering Mahan Hall West Point, NY 10996	2
Operational Test Command (OTC)	Aviation Test Directorate Ft. Hood, TX 76544	1
Dean, USMA	Office of the Dean Building 600 West Point, NY 10996	1
Defense Technical Information Center (DTIC)	ATTN: DTIC-O Defense Technical Information Center 8725 John J. Kingman Rd, Suite 0944 Fort Belvoir, VA 22060-6218	1
Department Head-DSE	Department of Systems Engineering Mahan Hall West Point, NY 10996	1
ORCEN	Department of Systems Engineering Mahan Hall West Point, NY 10996	5
ORCEN Director	Department of Systems Engineering Mahan Hall West Point, NY 10996	1
USMA Library	USMA Library Bldg 757 West Point, NY 10996	1